 RES HUMANITARIAE XVI, 2014, 126–143 ISSN 1822-7708

**Jolanta Kovalevskaitė** – humanitarinių mokslų daktarė, Vytauto Didžiojo universiteto jaunesnioji mokslo darbuotoja, lektorė.
*Moksliniai interesai:* tekstynų lingvistika, frazeologija, frazeografija, leksinė gramatika.
*El. paštas:* j.kovalevskaite@hmf.vdu.lt.

**Jolanta Kovalevskaitė:** Ph. D. in Humanities, junior researcher, lecturer of Vytautas Magnus University.
*Research interests:* corpus linguistics, phraseology, phraseography, lexical grammar.
*E-mail:* j.kovalevskaite@hmf.vdu.lt.

**Jolanta Kovalevskaitė**
*Vytauto Didžiojo universitetas*

# PHRASEME-TYPE AND PHRASEME-TOKEN: A CORPUS-DRIVEN EVIDENCE FOR MORPHOLOGICAL FLEXIBILITY OF PHRASEMES

**Anotacija**
Iš frazemų vartosenos galima ne tik nustatyti frazemos variantus, transformacijas ir modifikacijas, patikrinti, ar nėra vartosenos apribojimų, bet ir ištirti frazemos laisvumą. Žvalgomajame tyrime analizuoti pasirinkti lietuvių kalbos frazeologizmai su veiksmažodžiais, siekiant nustatyti, kokiomis formomis šie frazeologizmai įprastai vartojami (phraseme-type) ir kokiu dažnumu (phraseme-token). Palyginus duomenis apie frazeologizmų vartoseną *Dabartinės rašomosios lietuvių kalbos tekstyne*, *Frazeologijos žodyne* ir *Lietuvių kalbos daiktavardinių frazių žodyno* duomenų bazėje, matyti, kad iš tekstynų ir tekstynais paremtų žodynų išryškėja, kokiu laisvumu frazema pasižymi. Šia informacija papildžius lietuvių kalbos frazeologijos žodynus būtų galima sudaryti aprašus, geriau atskleidžiančius frazemos vartoseną.
PAGRINDINIAI ŽODŽIAI: frazema, tekstynas, frazemos forma (phraseme-type), frazemos formos pavartojimo dažnis (phraseme-token), variantiškumas, laisvumas.

**Abstract**
The usage of phrasemes evidences not only their variability, transformations and modifications, but also the most frequent forms of their realization (*phraseme-types*) and frequency (*phraseme-tokens*), i.e. phrasemes' flexibility. In this paper, selected Lithuanian idiomatic predicate phrasemes are analysed in the *Corpus of Contemporary Lithuanian Language*, in the *Phraseological Dictionary* and in the lexical database of the *Dictionary of Lithuanian*

*Phrases*. The results of comparison show that the corpus research can give rich evidence about the morphological flexibility of phrasemes. This information can help to improve representation of phrasemes in the phraseological dictionaries of Lithuanian, in order to make them more usage-based and more usage-oriented.
KEY WORDS: phraseme, corpus, phraseme-type, phraseme-token, variability, flexibility.

DOI: http://dx.doi.org/10.15181/ rh.v0i16.1016

## Introduction

Currently, the *fixedness* of a phraseme is understood as being of relative nature ("variable Stabilität" Stein 1995, 37, see Ridali 2012). Recent research on idioms has challenged the notion of non-compositionality of idioms and stressed their analysability and flexibility in discourse (for overview see Herold, Stathi 2007). Phraseology research based on corpora has shown that although being lexicalized holistic entities, idioms are more or less subject to variation. For example, Moon's (1998) findings showed that approximately 40% of 6776 fixed expressions in English have lexical variations or strongly institutionalized transformations. The study on German phrasemes (Ridali 2012) showed that out of 100 idioms, 41% are variable.

As Helju Ridali (2012) points out, the relative fixedness of idiomatic expressions can be seen through *variability and modification*[1]. In the typology of variability, presented by Harald Burger (1998, 25–27), the following cases are listed: a) substitutability by other morphological word forms (e.g., *seine Hand / seine Hände im Spiel haben*), b) substitutability by semantically similar words (e.g., *ein schiefes Gesicht machen / ziehen*), c) insertions of constituents or reducted forms (e.g., *sich etwas im Kalender anstreichen / sich etwas rot im Kalender anstreichen*), d) variability in word order (reversals) (e.g., *aussehen wie Milch und Blut / wie Milch und Blut aussehen*), and e) variability caused by valency (e.g., *sich die Schuhsohlen ablaufen nach etwas / um etwas zu bekommen*). Usually, the studies of the variants of idioms aim to detect the canonical form of an idiom, and to give the users of dictionaries or databases usage-oriented information,

[1] Variations (variants) are the variant forms of an individual expression with coincidentally matching meanings and with some common lexis (Moon 1998, 122). Modifications differ from variants in a way that variants are often institutionalized and listed in lexicographical resources. Modifications (also referred to as exploitations, e.g., Moon 1998, 170) are strongly related to a particular text and its author, which shows their occasional character (see Ridali for German 2012, 104; for Lithuanian see Butkutė 2010).

e.g., in *Idiomdatenbank*[2], the variants of a phraseme *das Blut ist in Wallung* ('someone is very excited') with its possible lexical substitutions and structural differencies are listed, e.g., *etw./jmd. bringt das Blut in Wallung, das Blut gerät in Wallung*, etc.

The discussion of the relative fixedness of idiomatic expressions has shown that not only the already mentioned phenomena of *variability* and *modification* are important, but also the issue of *impossible transformations of phraseme*. The latter is used as evidence for anomaly which helps to validate whether a specific phrase is a phraseme (Čermák 2007). In the entries of *The Dictionary of Czech Idioms*[3] (Čermák 2007, 162), negative categories, i.e., a set of individual anomalies of the idiom in question (e.g. impossible passivization, negation) are listed. Anomalies are especially important for computational phraseologists, but in order to detect the phrasemes automatically, a more detailed view of the features of phraseme behavior is necessary. Accordingly, phraseologists analyse the flexibility of phrasemes in discourse (e.g., discontinuous multi-word units, substituting constituents). In computational phraseology, the procedure of *lexical, morphosyntactic* and *syntactic variation* analysis (Heid 2008) covers the same phenomena described by two separate terms of *variability* and *impossible transformations* by simply answering the question "how much the given phraseme has in common with a normal phrase and to what extent the given phraseme is irregular (anomalous)?".

Although the variability of phrasemes includes cases when the constituents of phrasemes can be conjugated or declinated, most research focuses on the paradigmatic and syntagmatic features and questions whether the constituents of phrasemes can be replaced by other word forms or whether the same phraseme has extended and reducted variants. By observing the usage of phrasemes, we can detect how much phrasemes have in common with the normal (regularly built) phrases, i.e. to what extent a phraseme is flexible. Regularly built phrases are flexible with respect to their constituents which appear in a particular morphological form in discourse. As a majority of phrasemes do have "flexible" components which can be conjugated (declinated), each phraseme tends to be used in some particular form(s) in discourse.

[2] http://kollokationen.bbaw.de/

[3] Čermák F., Hronek J., Machač J. 1994: *Slovník české frazeologie a idiomatiky*. Academia Praha.

## 1. Pilot-study: Investigating the Evidence for Morphological Flexibility of Lithuanian Phrasemes

By observing the usage of phrasemes in discourse, one can see not only their lexical, morphosyntactic or syntactic variability, but also find out how the phraseme is used in context as a lexical item. Depending on their lexical meaning, one-word-lexemes are used in some particular word forms; the same works for a phraseme. To put it differently, if a particular phraseme is in some respect flexible, e.g., has a verbal component, then this component of the phraseme appears in a particular morphological form i.e., is conjugated. To give a concrete example, a Lithuanian phraseme "pakišti koją", meaning 'to put a spoke in smb's wheel', includes a verbal part "pakišti", which is used in *The Corpus of the Contemporary Lithuanian Language*[4] in an infinitive form ("pakišti"), 3rd person past ("pakišo"), 3rd person past frequentative ("pakišdavo"), 3rd person future ("pakiš"), and in conditional forms:

[1] , nes jis buvo įsitikinęs, jog baimė ir vėl **pakiš koją**, ir jis neišlaikys egzamino. Jam pasakiau: „Tu ti

[2] „Lietuvos rytui" sakė E. Gentvilas. Maskva **pakišo koją** ir parlamento valdybai 2002.4.23 Tadas Ignatavičiu

[3] us jis surimtėjo: „Kaip tik tai mums gali **pakišti koją** – juk latviai iš visų jėgų stengsis reabilituotis

[4] egimo projektais ir jų valdymu, dažnai **pakišančių koją** efektyviai projektų vadybai. Svarbiausia projektą

[5] kalai spaustų, lietus lyg tyčia man bandė **pakišti koją**. Nebijau aš jų – tamsių debesų ir įniršusio danga

[6] vaizduotę. Fantazijos dažnai praeityje **pakišdavo koją** skaistyklos suvokimui, teologai kalbėdavo apie ma

[7] įtariąs, kad šiuo atveju „autorei galėjo **pakišti koją** nepakankamai kryptingas ir tvirtas apsisprendimas

[8] lis noras tinkamai reprezentuotis, vis **pakišantis koją** laisvai kūrybai ir drąsai. Bernardinai.lt Popieži

[9] džeto spąstai Projekto įgyvendinimui gali **pakišti koją** ir prastai numatytas biudžetas. Pasak Simonos Buz

[10] kišti tauta, kaip Lisabonos sutarčiai jau **pakišo koją** airiai. Tai, kad referendumas gali neįvykti, šans

[11] ektams jie palaiko tik šalininkus ir visada **pakiš koją** oponentui. Geriausiu atveju rūmų intrigos kelia p

[4] http://tekstynas.vdu.lt/

This pilot study aims to describe the evidence of morphological flexibility of the selected Lithuanian phrasemes in the corpus and to investigate how the morphological flexibility is documented in the two phraseological dictionaries of Lithuanian. The study question of this paper determined the terminology: the form of usage in the corpus of a particular phrase is labelled as a *phraseme-type* (e.g., the above mentioned phraseme "pakišti koją" has a phraseme-type with the verb in 3rd person past "pakišo koją"), and the frequency of this particular phraseme-type as a *phraseme-token* (e.g., the type 3rd person past "pakišo koją" is used twice in the corpus)[5].

In other languages' phraseography, the topic of morphological features of the phraseme has already been discussed, mostly focusing on the fixedness rather than the flexibility of phrasemes. For instance, H. Burger (1998, 177) writes about the importance of 'morphosyntaktische Restriktionen' while discussing the lexicographic treatment of the phrasemes in German phraseological dictionaries. First, if a particular phraseme has the most typical usage form(s), it should be mentioned in the dictionary entry of this phraseme. Second, if a particular phraseme is realised only by some particular form, then this information has to be clearly presented for the user. R. Moon (1998, 7) uses the term 'lexicogrammatical fixedness' to refer to "lexicogrammatical defectiveness in units, for example, with preferred lexical realizations and often restrictions on aspect, mood, or voice". Terms used by H. Burger (1998) and R. Moon (1998) present the same position as in *The Dictionary of Czech Idioms,* i.e., to describe usage facts that show the fixedness of a phraseme. In this paper, we prefer to see the phenomena studied not as evidence for restriction, but rather as evidence for the **flexibility** of phrasemes (e. g., Wulff 2009). In the corpus-linguistically-based approach of V NP-constructions, adopted in S. Wulff (2009), three components of flexibility, each of which contains a specific set of variation parameters, were investigated: tree-syntactic flexibility, lexico-syntactic flexibility, and morphological flexibility. After the corpus analysis, 10 parameters for morphological flexibility were obtained. In our data the parameters of person, number, tense and aspect are observed.

For our pilot study, a subset of five 2- and 3-word phrasemes with the verbal part (predicate phrasemes) was used. The phrasemes that are idi-

[5] This terminological choice goes in one line with the terminology used in (Lithuanian) corpus research when word-form-type and word-form-token are discussed.

omatic, much more fixed than collocations, and can be labelled as transparent metaphors (Moon 1998, 19) were investigated. In the first part of the pilot study, a comparison of the Lithuanian idiom dictionary (*Phraseological Dictionary* by Paulauskas 2001) and *The Corpus of the Contemporary Lithuanian Language* was performed, in order to compare the information output about the phraseme flexibility in both of these resources (see section 1.1.). In the second part of the pilot study (section 1.2.), the focus was on the presentation of the same subset of phrasemes in the corpus-based database of *The Dictionary of Lithuanian Phrases* (Rimkutė, Bielinskienė, Kovalevskaitė 2012). After the discussion of the results, we conclude by stressing the importance and need for more studies of the flexibility of phrasemes and better representation of the phenomenon in the Lithuanian phraseography.

### 1.1. Dictionary vs. Corpus

As already mentioned, by studying the usage of a particular phraseme we can see not only anomalies or variants, but also the typical usage of the phraseme, i.e., the most frequent forms of its (constituents) realization in corpus.

Selected predicate phrasemes which were investigated here are included in *Phraseological Dictionary* (Paulauskas 2001) and found in *The Corpus of the Contemporary Lithuanian Language. The Corpus of the Contemporary Lithuanian Language* is a general monolingual corpus of almost 200 million words from texts drawn from a variety of genres and representing modern written Lithuanian (1992–2006).

*Phraseological Dictionary* (Paulauskas 2001) is an idiomatic dictionary, representing idioms and phraseologisms of Lithuanian. Information concerning the frequent types of phrasemes is not explicitly explained in this dictionary. The usage-oriented information provides the usage labels (e.g., ironical, formal, pejorative, etc.) and selected examples (illustrations) of the data. As the majority of the Lithuanian phraseological dictionaries were compiled when there were no Lithuanian language corpora, examples were selected from literature and other sources (e.g., *The Contemporary Lithuanian Dictionary, The Academic Lithuanian Dictionary*), thus the frequency information is not available.

First, the forms of the verbal part of the selected phrasemes in the dictionary and the forms of the flexible part of the selected phrasemes in the corpus were analysed. By comparing the information given about the usage of phrasemes in the dictionary with the data from the corpus, it has been noticed that the dictionary does not always give a clear indication which verb form(s) included in the particular phraseme is (are) used in context. For example, for the phraseme "į galvą šauti" ('to get/take it into one's head (to do something)') there are 5 forms of the verb "šauti" given in the dictionary: "šauti" (inf), "šovė" (3rd person, past), "šauna" (3rd person, present), "šaus" (3rd person, future), "šovęs" (3rd person, past participle). From the given examples, it is not clear whether some forms are more frequent than others (except from "šovė", which is used 5 times in illustrations).

By examining the usage of a phraseme in the corpus, we can find not only word forms (phraseme-types), but also the frequency of a particular phraseme-type (phraseme-tokens), thus, we can better realize how the given phraseme appears naturally in discourse and how it looks like formally (see Table 1). For example, 3rd person past "šovė" (67 phraseme-tokens), 3rd person present "šauna" (27 phraseme-tokens), infinitive "šauti" (8 phraseme-tokens), 3rd person conditional "šautų" (8 phraseme-tokens), negated 3rd person present and past "nešauna" (7 phraseme-tokens), "nešovė" (6 phraseme-tokens), 3rd person future "šaus" (4 phraseme-tokens). Table 1 summarizes the results of the comparison of the data from the corpus and from the dictionary.

From the comparison of the frequency of the phraseme-types in the dictionary and in the corpus, several observations can be made. First, in the dictionary, only illustrations are given with no information which phraseme-type is more frequent. As there are cases when phraseme-type listed in the dictionary does not appear in the corpus, frequency is important in order to identify the mostly used phraseme-types. However, quite often, the most frequently used phraseme-type in the corpus is given as the first in the dictionary as well (see phraseme-types *duoda garo, šovė į galvą, eina iš proto*). On the other hand, there are cases when frequently used phraseme-type is not listed in the dictionary examples at all (e.g., *kabo ant plauko*). In the dictionary, often only one or two forms are used several times, whereas in the corpus, one can find more examples, and, accordingly, see more differences in the usage and differentiate between those phraseme-types that are more frequent than others.

Table 1

**Phraseme-types and frequency of the phrasemes studied in the corpora (*phraseme-tokens*)**

| **Phraseme and its meaning** | **Phraseme-types, sorted by the morphological features of the verbal part of phraseme** | **Phraseme-types' occurence in dictionary** | **Phraseme-types' occurence in corpus** |
|---|---|---|---|
| **Duoti garo** 'to read the riot act' | *Duoti* (infinitive) *garo* | 1 | – |
| | *Duoda* (3rd person, present) *garo* | 3 | 7 |
| | *Davė* (3rd person, past) *garo* | 3 | 1 |
| | *Davėm* (1st person, past, plural) *garo* | 1 | – |
| | *Duos* (3rd person, future) *garo* | 1 | 4 |
| | *Duosim* (1st person, future, plural) *garo* | 1 | – |
| | *Duok* (2nd person, imperative) *garo* | 1 | 1 |
| **Pakišti koją** 'to put a spoke in smb's weel' | *Pakišti* (infinitive) *koją* | 3 | – |
| | *Pakiša* (3rd person, present) *koją* | – | 6 |
| | *Pakišo* (3rd person, past) *koją* | 1 | 10 |
| | *Pakišdavo* (3rd person, past frequentative) *koją* | – | 3 |
| | *Pakiš* (3rd person, future) *koją* | – | 3 |
| **Šauti į galvą** 'to get it into one's head (to do something)' | *Šauti* (infinitive) *į galvą* | 1 | 8 |
| | *Šauna* (3rd person, present) *į galvą* | 1 | 27 |
| | *Nešauna* (3rd person, present, negated) *į galvą* | – | 7 |
| | *Šovė* (3rd person, past) *į galvą* | 5 | 67 |
| | *Nešovė* (3rd person, past, negated) *į galvą* | – | 6 |
| | *Šaudavo* (3rd person, past frequentative) *į galvą* | – | 3 |
| | *Šovęs* (3rd person, conditional, past, masculine) *į galvą* | 1 | – |
| | *Šovusi* (3rd person, conditional, past, feminine) *į galvą* | – | 3 |
| | *Šovę* (2rd person, conditional, past, plural) *į galvą* | – | 3 |
| | *Šaus* (3rd person, future) *į galvą* | 1 | 4 |
| | *Šautų* (3rd person, conditional) *į galvą* | – | 8 |

| **Phraseme and its meaning** | **Phraseme-types, sorted by the morphological features of the verbal part of phraseme** | **Phraseme-types' occurence in dictionary** | **Phraseme-types' occurence in corpus** |
|---|---|---|---|
| **Kaboti ant plauko** 'cliffhang' | *Kaboti* (infinitive) *ant plauko* | 1 | 1 |
| | *Kabo* (3rd person, present) *ant plauko* | – | 7 |
| | *Kaba* (3rd person, present) *ant plauko* | 1 | – |
| | *Kabojo* (3rd person, past) *ant plauko* | 2 | 2 |
| **Eiti iš proto** 'to be out of one's mind' | *Eiti* (infinitive) *iš proto* | 1 | 4 |
| | *Eina* (3rd person, present) *iš proto* | 5 | 16 |
| | *Einu* (1st person, present) *iš proto* | 1 | 10 |
| | *Eini* (2nd person, present) *iš proto* | 1 | 5 |
| | *Neina* (3rd person, present, negated) *iš proto* | – | 3 |
| | *Ėjo* (3rd person, past) *iš proto* | 1 | 7 |
| | *Eis* (3rd person, future) *iš proto* | 1 | – |

Of course, if a phraseme is not frequent, we cannot find strong evidence about its usage from the corpus. Other studies have evidenced that the frequency of phrasemes highly depends on the text type (see Biber 2009; Pivovarova, Yagunova 2010). Thus, for studying the types and tokens of a particular phraseme or phraseme group(s), the corpus should be chosen properly, having in mind the research objectives. Going back to the findings of this research, it has to be stressed that in such general and relatively large corpus as *The Corpus of the Contemporary Lithuanian Language*, we can expect good preconditions to study frequency even of those phrases that are used not frequently, except of archaic phrasemes, or those which are lesser-used in contemporary Lithuanian, as well as those which are typical for spoken Lithuanian.

The analysis of the morphological features of the verb forms in phraseme-types (Table 1) shows that although infinitive is given in the dictionary examples, it is rather seldom (see *duoti garo, pakišti kojį*) in discourse. From language learners' perspective, a following question could be asked: if phrasemes are described as language units which are (more or less) of restricted nature, then, probably, it would be worth informing in which form a particular phraseme appears most often, instead of only giving infinitive-lemma as it is usual with one-word language units.

In the corpus data, a greater morphological richness can be seen (e.g., phraseme-types of *pakišti koją pakišti, šauti į galvą šauti, eiti iš proto*). A number of cases have been found when the same phraseme-types are used both in the corpus and in the dictionary. However, as noticed above, these forms differ in their frequency (e.g., *pakišo* (3rd person, past) *kojį, šauna* (3rd person, present) *į galvą, šovė* (3rd person, past) *į galvą, eina* (3rd person, present) *iš proto*). As form relates to the meaning very closely, a more detailed and comprehensive study could ask a question to what extent the flexibility of the phraseme can be important in making decisions about the motivation (and/or compositionality) of the phraseme.

A deeper analysis of how the form and meaning are connected could give stronger usage-based arguments for better entries in the phraseological learner-oriented dictionaries[6]. If phrasemes' meaning is taught together with its form, better results can be achieved as learner associates what the phraseme means and how this meaning is seen in the form of phraseme (semantic motivation). For example, a phraseme *eiti iš proto* is often realized in the corpus by present forms: *eina, einu, eini*. The phraseme describes a situation where feeling is expressed, which can be connected with "now and here". Negated forms of phraseme-type *nešovė į galvą* are also related to the communicative needs, when the speaker wants to express his/her disappointment that he/she could not predict some event just because he/she has not thought about it. These ideas go in line with D. Siepmann's (2008, 199) perspective, who, in the discussion of one type of phrasemes, transparent collocations, points out that "collocations are inextricably linked with, and often restricted to, a particular topic area or situation type through what may be described as neuronal assemblies, that is, the repeated association of lexical units or semantic-pragmatic features with a situational or syntagmatic context". Being aware of this situational context, a learner can deal with the acquisition of phrasemes more effectively.

### 1.2. Corpus-based Dictionary vs. Corpus

After examining the selected phrase set in the corpus and in the dictionary, the data from *The Dictionary of Lithuanian Phrases* is compared against the corpus data.

[6] The importance of presenting usage information in phrasemes' teaching is illustrated in the study of Boers, Eyckmans, Stengers (2007), where teaching processed associating an idiom with its etymology.

*The Dictionary of Lithuanian Phrases* (Rimkutė, Bielinskienė, Kovalevskaitė 2012) is a corpus-based electronic dictionary and electronically published lexical database[7]. The dictionary consists of: 1) the plain list of fixed phrases, alphabetically filtered by the first word of the phrase; 2) the online database, based on morphologically annotated phrase list. The search in the online database can be carried out by using: a) main options (search by first word, word form, a part of the word form, a particular phrase or its part); b) advanced options (search by morphological features of the words of the phrase). The dictionary includes phrases of different length, which were automatically extracted from the first version of *The Corpus of Contemporary Lithuanian* consisting of 100 mln running words and texts written in 1991–2002[8].

For the automatical extraction of the phrases, first, the list of statistically significant collocational chains was generated. For the extraction of chains, a new method, *Gravity Counts,* was developed (Daudaravičius, Marcinkevičienė 2004, 330). *Gravity Counts* helps to evaluate the combinability of two words according to individual word frequencies, pair frequencies or the number of different words in the selected 3 word-span. The method allowed to detect the collocational chains without using a list of the previously selected node-words. As a result, many text fragments of varying length were extracted (e.g., 2-word-10-word) (for the information about the manual procedures and how the collocational chains are transformed into well-formed phrases see Marcinkevičienė, Grigonytė 2005; Rimkutė, Bielinskienė, Kovalevskaitė 2012). The extracted phrases contain collocating grammatical forms presented in their natural word order and in the form they appear in the corpus. As Lithuanian is a highly morphologically rich language, it is an advantage that in *The Dictionary of Lithuanian Phrases,* the phrases are not lemmatized, but given in the form they appear in the corpus, e.g., *kabo ant plauko*, *kabojo ant plauko*, *pakibo ant plauko*.

In The *Dictionary of Lithuanian Phrases,* several phraseme-types of the same phraseme can be found. Accordingly, we can compare the list of several phraseme-types for each phrase detected by Gravity Counts method

[7] http://tekstynas.vdu.lt/page.xhtml;jsessionid=953769EE98B4A3426313B33FBD020A2B?id=dictionary-db

[8] It is the same corpus we use in this study.

automatically as well as phraseme-types manually found in the corpus. The comparison can be seen in Table 2.

Table 2

**The phraseme-types in *The Dictionary of Lithuanian Phrases* and in the corpus[9]**

| **Phraseme** | **Phraseme-types in *The Dictionary of Lithuanian Phrases*** | **Phraseme-types in the corpus** |
|---|---|---|
| **Duoti garo**<br>'to read the riot act' | DUOKIM GARO | duoda garo (7),<br>duos garo (4),<br>davė garo,<br>duodama garo,<br>duok garo |
| **Pakišti koją**<br>'to put a spoke in smb's weel' | **KOJĄ PAKIŠO...**<br>**PAKIŠO KOJĄ**<br>PAKIŠTI KOJĄ | **koją pakišo (10)**,<br>pakiša koją (6),<br>koją pakišdavo (3),<br>koją pakiš (3) |
| **Šauti į galvą**<br>'to get it into one's head (to do something)' | ŠAUNA Į GALVĄ<br>ŠAUS Į GALVĄ<br>ŠOVĖ Į GALVĄ | šovė į galvą (67),<br>šauna į galvą (27),<br>šauti į galvą (8),<br>šautų į galvą (8),<br>nešauna į galvą (7),<br>nešovė į galvą (6),<br>šaus į galvą (4),<br>šaudavo į galvą (3),<br>šovusi į galvą (3),<br>į galvą šovę (3) |
| **Kaboti ant plauko**<br>'cliffhang' | **KABO ANT PLAUKO**<br>**KABOJO ANT PLAUKO**<br>PAKIBO ANT PLAUKO | **kabo ant plauko (7)**,<br>**kabojo ant plauko (2)**,<br>kaboti ant plauko |
| **Eiti iš proto**<br>'to be out of one's mind' | **eina**; kraustosi **IŠ PROTO** | **iš proto eina (16)**,<br>einu iš proto (10),<br>ėjo iš proto (7),<br>eini iš proto (5),<br>eiti iš proto (4),<br>neina iš proto (3) |

[9] For the phraseme-types found in the corpus, their frequency is given in brackets. In bold those phraseme-types that overlap in the dictionary and in the corpus are given.

First, it is seen that the most frequent phraseme-types detected automatically are those with the highest frequency in the corpus (e.g., *kojq pakišo* – 10 occurences, *šovė į galvą* – 67, *šauna į galvą* – 27, *kabo ant plauko* – 7, *iš proto eina* – 16 occurences). Therefore, the corpus-based *Dictionary of Lithuanian Phrases* can be seen as a useful resource to evidence the morphological flexibility of Lithuanian phrasemes. Of course, it is an advantage of the Gravity Counts method that phraseme-tokens can be automatically extracted. However, this resourse does not help to answer all questions necessary for the documentation of the morphological flexibility of the phrasemes in the phraseological dictionaries.

If some phraseme-type is included in the database, it means that the words building this phraseme-type are collocating strongly and are used as one unit in the discourse. Thus, we get the list of the most frequently used phraseme-types of a particular phrase, but we do not get the information about the frequency of each of these phraseme-types. Another problematic point is that concordance lines can be sorted and idiomatic usage from non-idiomatic can be separated only manually. Exactly for this reason, in the Dictionary of Lithuanian Phrases, phraseme-type šaus į galvą, which occurs only 4 times in the corpus, was found. It is very likely that this phrase is not a type of the fixed phrase šauti į galvą which means 'to get it into one's head (to do something), but it is a regular phrase meaning "shoot into head"10. There is another similar example: the phraseme-type pakišti koją was not detected by manual analysis of the concordance of the fixed phrase "pakišti koją" ('to put a spoke in smb's weel'), because it was expression of a literal meaning as in "he fell over smb's stretched leg". A reason, why the phrase "Duokim garo" was detected automatically, but does not appear between phraseme-types of the fixed phrase "duoti garo" ('to read the riot act') is that this particular phrase is used as a title of a popular TV show11.

Although more data should be examined, it is already seen that the database of *The Dictionary of Lithuanian Phrases* can be used as a resource to identify and to describe frequently used phraseme-types. Nonetheless,

[10] Although the phraseme-type šovė į galvą is included in the database and detected during our corpus analysis, there is still a real possibility that a part from all instances of this phraseme-type in the database is not the idiomatic usage.

[11] In manual corpus analysis, names, titles etc., are excluded.

for the phrasemes which can have literal meaning, manual (or additional automatic) concordance analysis has to be done in order to identify only idiomatic usage. For studying the most frequent patterns of language, a large amount of corpus data has to be examined. Therefore, it can be seen as an advantage that by applying statistical tools, phraseographers can get prepared data (see, for example, *UWV-Analysemodell* in Steyer, Brunner (2009)), and to produce lexical resources where the real usage of phraseological items is documented.

## Conclusions

The phenomenon of variability of phrasemes is widely discussed in the literature of phraseology. However, when analyzing the usage of phrasemes of different types, especially in corpora, together with the phenomena of variability, modification and transformation, clear evidence for phraseme's flexibility is also found, e.g., the most frequent forms of phraseme's realization in usage. In this article, we analyse the morphological flexibility of selected phrasemes, using two labels: "phraseme-type" and "phraseme-token". The information about this lesser discussed aspect of phraseme's usage is especially relevant for such morphologically rich language as Lithuanian: the results show that depending on a phraseme, it can appear in 3 to 10 different phraseme-types in the corpus.

The comparison of the data from the *Phraseological Dictionary*, the *Corpus of Contemporary Lithuanian Language*, and *The Dictionary of Lithuanian Phrases* has shown that phraseme-types are simply listed in the dictionary in examples, i.e., dictionary does not include information about phraseme-type frequency. Often, the same phraseme-type appears in several examples, and, probably, this fact can indicate for the user the higher frequency of this phraseme-type. Often, the most frequently used phraseme-type in the corpus is given as first in the dictionary as well, but it is not the direct indication of the most frequent phraseme-type. There are phraseme-types in the dictionary, which do not appear in the corpus at all. When the same phraseme-types are used both in the corpus and in the dictionary, it shows that examples included in the *Phraseological Dictionary* (Paulauskas 2001) represent the earlier usage which did not change in contemporary written Lithuanian. In the corpus data, a greater morpho-

logical richness can be seen. Thus, in order to identify the most frequent phraseme-types, we have to investigate phraseme-tokens in the corpus.

Automatically extracted data in *The Dictionary of Lithuanian Phrases* can help phraseologists to analyse the phenomenon of the morphological flexibility: the phraseme-tokens are automatically extracted, and, as the pilot study shows, the quality of the data is good. On the other hand, we do not see the information about the frequency of each phraseme-type, and if the phrase can be used in literal and idiomatic sense, then manual (or additional automatic) concordance analysis has to be carried out in order to distinguish only phraseme-types from idiomatic usage. Thus, although the possibilities to study the flexibility of phrasemes using corpus tools are considerably improving, we have to know the possible shortcommings.

It was shown that corpus analysis gives evidence about the phraseme's morphological flexibility. This pilot study is an attempt to underline the importance and need for more studies of the flexibility of phrasemes and better representation of the phenomenon in the Lithuanian phraseography. Not only the restrictions, but also flexibility is important for lexicographers and dictionary users, because more detailed information concerning phraseme's usage can help to separate flexible phrasemes from fully fixed and frozen phrasemes, i.e., to give a clearer idea of the diversity even of such subset of phrasemes that are represented as fixed. In the usage-based and user-oriented phraseography, more attention has to be given to document how the phraseme is used.

As form relates to meaning very closely, a more detailed and comprehensive study could ask a question to what extent the flexibility of a phraseme can be important for making decisions about the motivation (and/or compositionality) of the phraseme. To study morphological flexibility of the Lithuanian phrasemes deeper, more data has to be considered in order to describe various aspects of formal flexibility and to give a more systhematic picture of the phenomenon.

## References

Biber 2009 – Douglas Biber. A Corpus-driven Approach to Formulaic Language in English. Multi-word Patterns in Speech and Writing. *International Journal of Corpus Linguistics 14 (3)*, 275–311.

Boers, Eyckmans, Stengers 2007 – Frank Boers, June Eyckmans, Hélène Stengers. Presenting Figurative Idioms with a Touch of Etymology: More than Mere Mnemonics? *Language Teaching Research 11 (1)*, 43–62.

Burger 1998 – Harald Burger. *Phraselogie. Eine Einführung am Beispiel des Deutschen*. Berlin: Erich Schmidt.

Butkutė 2010 – Laura Butkutė. *Okazionaliųjų frazeologizmų stilistinės išgalės publicistiniame stiliuje*. Daktaro disertacija. Vilnius: Vilniaus universiteto leidykla.

Čermák 2007 – František Čermák. Substance of Idioms: Perennial Problems, Lack of Data or Theory? (first published in: *International Journal of Lexicography 14*, 2001, 1–20). *Czech and General Phraseology*. Praha: Carolinum, 149–166.

Čermák, Hronek, Machač 1994 – Čermák F., Hronek J., Machač J. *Slovník české frazeologie a idiomatiky*. Academia Praha.

Daudaravičius, Marcinkevičienė 2004 – Vidas Daudaravičius, Rūta Marcinkevičienė. Gravity Counts for the Boundaries of Collocations. *International Journal of Corpus Linguistics 9 (2)*, 321–348.

Heid 2008 – Ulrich Heid. Computational Phraseology: and Overview. *Phraseology. An Interdisciplinary Perspective* (eds. Sylviane Granger, Fanny Meunier). Amsterdam: John Benjamins, 337–360.

Herold, Stathi 2007 – Axel Herold, Katerina Stathi 2007: Measuring Syntagmatic Fixedness of Multi-Word Expressions. *Proceedings of the Corpus Linguistics Conference* (eds. Matthew Davies, Paul Rayson, Susan Hunston, Pernilla Danielsson). Accessed August 30, 2013, http://ucrel.lancs.ac.uk/publications/CL2007/paper/80_Paper.pdf.

Marcinkevičienė, Grigonytė 2005 – Rūta Marcinkevičienė, Gintarė Grigonytė. Lexicogrammatical Patterns of Lithuanian Phrases. *Proceedings of the Second Baltic Conference on Human Language Technologies*, 299–305.

Moon 1998 – Rosamund Moon. *Fixed Expressions and Idioms in English. A corpus-based approach*. Oxford: Clarendon Press.

Paulauskas 2001 – Jonas Paulauskas. *Frazeologijos žodynas*. Vilnius: Lietuvių kalbos institutas.

Pivovarova, Yagunova 2010 – Lidia Pivovarova, Elena Yagunova. Collocation Extraction and Text Analysis: *Different Types of Collocations and Different Genres. SLTC Workshop on Compound and Multiword Expressions*. Accessed August 30, 2013, http://www.ida.liu.se/~sarst/compound-ws/papers/pivovarova_yagunova.pdf.

Ridali 2012 – Helju Ridali. Wie fest sind feste Wendungen? Eine Untersuchung zur morphosyntaktischen Variabilität in deutschen Phraseologismen. *Triangulum. Germanistisches Jahrbuch 2011 für Estland, Lettland und Litauen*. Vilnius: Verlag der Kunstakademie Vilnius, 101–115.

Rimkutė, Bielinskienė, Kovalevskaitė 2012 – *Lietuvių kalbos daiktavardinių frazių žodynas* (red. Erika Rimkutė, Agnė Bielinskienė, Jolanta Kovalevskaitė). Kaunas: Vytauto Didžiojo universiteto leidykla. Accessed August 30, 2013, http://donelaitis.vdu.lt/lkk/pdf/daikt_fr.pdf.

Siepmann 2008 – Dirk Siepmann. Phraseology in Learner's Dictionaries. What, Where and How? *Phraseology in Foreign Language Learning and Teaching* (eds. Fanny Meunier, Sylviane Granger). Amsterdam/Philadelphia: John Benjamins, 185–202.

Steyer, Brunner 2009 – Kathrin Steyer, Annelen Brunner. Das *UWV*-Analysemodell. Eine korpusgesteuerte Methode zur linguistischen Systematisierung von Wortverbindungen.

*OPAL* – Online publizierte Arbeiten zur Linguistik 1/2009: Accessed August 30, 2013, http://pub.ids-mannheim.de/laufend/opal/.
Wulff 2009 – Stefanie Wulff. Converging Evidence from Corpus and Experimental Data to Capture Idiomaticity. *Corpus Linguistics and Linguistic Theory 5 (1)*, 131–159.

**Jolanta Kovalevskaitė**

## FRAZEMŲ VARTOSENOS YPATYBĖS TEKSTYNUOSE: FRAZEMOS FORMA IR FRAZEMOS FORMOS PAVARTOJIMO DAŽNIS

Santrauka

Naujausi tekstynų analize paremti frazeologijos tyrimai rodo, kad stabiliesiems junginiams (frazemoms) būdingas nevienodas sustabarėjimo laipsnis ir kad dauguma jų pasižymi mažesniu ar didesniu variantiškumu (Moon 1998, Ridali 2012). Tos pačios frazemos variantai ir nuo autoriaus itin priklausančios įprastų, paplitusių frazemų neįprastos modifikacijos yra požymiai, kuriais remiamasi įrodant, kad frazemos nėra visiškai sustabarėjusios. Įvairūs formalieji, morfosintaksiniai, frazemų vartosenos apribojimai – atvirkščiai – taikomi vertinant frazemų sustabarėjimo laipsnį. Frazeografijoje stengiamasi atskleisti variantiškumą arba nurodyti frazemos sustabarėjimo požymius (žr. *lexicogrammatical deffectiveness* (Moon 1998), *morphosyntaktische Restriktionen* (Burger 1998)), tačiau taip pati frazemos vartosena yra atskleidžiama tik iš dalies.

Analizuojant frazemų vartoseną tekstyne, išryškėja ne tik frazemų variantai, bet ir frazemoms įprastos, dažniausios formos. Stabilieji junginiai yra leksiniai vienetai, kurie, kaip ir vienažodžiai leksiniai vienetai, realizuojami tekste įgyja vienokią ar kitokią raišką. Vadinasi, tyrinėdami, kaip frazema vartojama, galime nustatyti pačias dažniausias frazemos formas (*phraseme-type*) ir tų formų pavartojimo dažnį (*phraseme-token*). Žvalgomajame tyrime analizuoti pasirinkti lietuvių kalbos frazeologizmai su veiksmažodžiu (*pakišti kojį, kaboti ant plauko, eiti iš proto* ir kt.), aprašyti Jono Paulausko *Frazeologijos žodyne* (2001) ir vartojami *Dabartinės rašomosios lietuvių kalbos tekstyne*. Pirmiausia tyrinėta, kiek frazeologizmų aprašas žodyne atitinka tų frazeologizmų vartoseną tekstyne, vėliau tekstyno duomenys palyginti su žodyno, paremto tekstynu, duomenimis.

Palyginus duomenis apie šių frazeologizmų vartoseną *Dabartinės rašomosios lietuvių kalbos tekstyne* ir *Frazeologijos žodyne*, pastebėta, kad žodyno iliustracijose pateikiami tik pavyzdžiai su skirtingomis tam tikros frazemos formomis (pvz., *į galvą šauti, į galvą šovė, į galvą šauna...*), o iš tekstyno galime matyti ne tik kurios iš šių formų yra vartojamos, bet ir koks jų pavartojimo dažnumas (pvz., *į galvą šovė* (67 pavartojimo atvejai), *į galvą šauna* (27). Tik iš tekstyno pavyzdžių išryškėja šio frazeologizmo forma *nešauna į galvą* (7) ir *nešovė į galvą* (6), kuri, jeigu atsižvelgtume į šio junginio reikšmę, gali būti šiam frazeologizmui labai būdinga.

*Lietuvių kalbos daiktavardinių frazių žodyno* duomenų bazėje pateikiami duomenys yra surinkti pritaikius kolokacijų atpažinimo statistinį metodą *Gravity Counts*, kuris leidžia ne tik nustatyti įvairaus ilgio junginius, bet ir išsaugoti informaciją apie tuos junginius sudarančių žodžių formą. Palyginus, kaip tyrinėjami frazeologizmai pateikiami šioje bazėje, paaiškėjo, kad dažniausiai jie atpažinti būtent ta forma, kuria tekstyne pavartoti dažniausiai (pvz., *kabo ant plauko, kabojo ant plauko*). Vis dėlto, nors šio žodyno duomenys pateikia informacijos apie analizuotų frazeologizmų formas, tačiau be duomenų apie tų formų pavartojimo dažnumą. Be to, analizė parodė, kad duomenų tikslumas menkesnis, jeigu analizuojamas junginys gali būti vartojamas ir kaip frazeologizmas, ir kaip laisvasis žodžių junginys, t. y. tiesiogine reikšme (pvz., *pakišo kojų, šauti į galvą*).

Duomenys apie frazemų laisvumą frazeografijai svarbūs ir teoriniu, ir praktiniu požiūriu – siekiant aiškiau atskirti labiau sustabarėjusias frazemas nuo mažiau sustabarėjusių (pasižyminčių didesniu laisvumu) ir išsamiau aprašyti įvairius frazemų tipus, arba rengiant tekstynais paremtus frazeologijos žodynus ir leksines bazes, kurios itin praverstų lietuviškosios frazeologijos besimokantiems asmenims, nes tokiuose žodynuose būtų geriau atskleistos frazemų vartosenos ypatybės.